

FusionKG-Rural: Fusão Multimodal de Dados Cadastrais, Fiscais e Ambientais para Associação de Imóveis Rurais

Roberval G. Mariano¹, Vânia P. Vidal²

¹Departamento de Ciência da Computação – Universidade Federal do Ceará (UFC)
Caixa Postal 60.020 – Fortaleza – CE – Brazil

{mariano.rgm.2015,vaniap}@gmail.com

Abstract. Integrating rural data from CAFIR, SNCR and SICAR/CAR remains challenging due to the absence of universal identifiers and the semantic, spatial and cadastral heterogeneity of these datasets. This paper presents FusionKG-Rural, a multimodal data fusion architecture based on Knowledge Graphs and specialized agents for automatic identification of records representing the same rural property. The approach combines semantic, geospatial, cadastral and fiscal evidence through a Dempster-Shafer-inspired fusion mechanism. Results indicate that combining multiple sources of evidence reduces ambiguities and improves entity matching quality, supporting territorial governance, tax inspection and environmental monitoring applications.

Resumo. A integração de dados rurais provenientes do CAFIR, SNCR e SICAR/CAR é dificultada pela ausência de identificadores universais e pela heterogeneidade semântica, espacial e cadastral das bases. Este trabalho apresenta o FusionKG-Rural, uma arquitetura de fusão multimodal baseada em Grafos de Conhecimento e agentes especializados para identificação automática de registros que representam o mesmo imóvel rural. A solução combina evidências semânticas, geoespaciais, cadastrais e fiscais por meio de um mecanismo inspirado na Teoria de Dempster-Shafer. Os resultados indicam que a combinação de múltiplas evidências reduz ambiguidades e melhora a qualidade da associação entre registros heterogêneos, fornecendo suporte para aplicações de governança territorial, fiscalização tributária e monitoramento ambiental.

1. Introdução

A administração territorial brasileira depende de múltiplos sistemas de informação mantidos por diferentes órgãos governamentais. Entre os principais destacam-se o Cadastro de Imóveis Rurais da Receita Federal (CAFIR), o Sistema Nacional de Cadastro Rural (SNCR) e o Sistema Nacional de Cadastro Ambiental Rural (SICAR/CAR).

Embora essas bases descrevam um mesmo objeto do mundo real, o imóvel rural e seus registros são produzidos independentemente, utilizando estruturas de dados, identificadores e critérios de atualização distintos.

Esse cenário produz problemas clássicos de fusão de dados: duplicidades, inconsistências, incompletudes e conflitos semânticos.

Este trabalho propõe o FusionKG-Rural, uma arquitetura baseada em Grafos de Conhecimento e fusão de evidências para associação automática de imóveis rurais presentes em bases heterogêneas.

A Teoria de Dempster-Shafer foi escolhida por permitir representar explicitamente incerteza e conflito entre evidências provenientes de fontes heterogêneas. Modelos probabilísticos tradicionais exigem probabilidades completas, enquanto Dempster-Shafer permite representar conhecimento parcial.

As contribuições científicas deste trabalho são: (C1) Um modelo de fusão multimodal para associação de imóveis rurais integrando evidências fiscais, cadastrais, geoespaciais e semânticas; (C2) Uma modelagem baseada em Grafos de Conhecimento para representação unificada dos registros CAFIR, SNCR e SICAR/CAR; (C3) Um mecanismo de combinação de evidências baseado em Dempster-Shafer capaz de representar explicitamente conflito e incerteza entre fontes heterogêneas; (C4) Uma avaliação em larga escala envolvendo mais de 13 milhões de registros rurais.

2. Trabalhos Relacionados

A fusão de dados constitui uma área consolidada de pesquisa voltada à integração de informações provenientes de múltiplas fontes heterogêneas. Hall e Llinas (1997) definem Data Fusion como o processo de combinação de dados oriundos de diferentes sensores ou sistemas com o objetivo de produzir conhecimento mais consistente e confiável do que aquele obtido individualmente por cada fonte. Posteriormente, Khaleghi et al. (2013) ampliaram essa visão ao apresentar um panorama abrangente dos desafios relacionados à heterogeneidade, inconsistência e incerteza em ambientes de fusão multimodal.

No contexto da resolução de entidades, Christen (2012) sistematizou técnicas de Record Linkage e Entity Resolution para identificação de registros equivalentes em bases de dados heterogêneas. Mais recentemente, Zhao et al. (2023) destacam que o problema de alinhamento de entidades em Grafos de Conhecimento tornou-se um dos principais desafios da integração semântica, especialmente quando diferentes fontes descrevem o mesmo objeto do mundo real utilizando esquemas e identificadores distintos.

A utilização de Grafos de Conhecimento como mecanismo de integração semântica tem crescido significativamente nos últimos anos. Hogan et al. (2021) apresentam uma visão abrangente sobre construção, representação e exploração de Knowledge Graphs em larga escala. Complementarmente, Paulheim (2023) discute técnicas de refinamento, enriquecimento e consolidação de grafos, destacando o papel da fusão de entidades e da resolução de inconsistências na melhoria da qualidade dos dados integrados.

A construção de Grafos de Conhecimento a partir de múltiplas fontes introduz novos desafios relacionados à fusão de informações conflitantes. Nesse contexto, Li et al. (2024) apresentam um levantamento recente sobre técnicas de Data Fusion aplicadas à construção de Knowledge Graphs, enfatizando a necessidade de mecanismos capazes de combinar evidências provenientes de fontes com diferentes níveis de confiabilidade.

Entre os modelos de fusão de evidências, a Teoria de Dempster-Shafer permanece como uma das abordagens mais utilizadas para representação explícita de

crença, plausibilidade e incerteza. Conforme proposto por Shafer (1976), a combinação de evidências permite integrar informações parcialmente conflitantes sem exigir distribuições probabilísticas completas, característica particularmente relevante em cenários de integração cadastral, fiscal e ambiental.

Embora existam avanços significativos em Data Fusion, Entity Resolution e Knowledge Graph Fusion, observa-se uma lacuna relacionada à associação de imóveis rurais utilizando simultaneamente evidências fiscais, cadastrais, geoespaciais e ambientais. O presente trabalho busca preencher essa lacuna por meio de uma arquitetura multimodal baseada em Grafos de Conhecimento e fusão de evidências aplicada às bases CAFIR, SNCR e SICAR/CAR.

3 Formalização do FusionKG-Rural

Considere três bases de dados distintas: $B = \{CAFIR, SNCR, CAR\}$. Cada registro pertencente a uma dessas bases é representado como uma entidade em um Grafo de Conhecimento Rural: $G = (V, E)$, onde: V representa o conjunto de entidades e E representa o conjunto de relacionamentos. Os vértices podem representar imóveis, proprietários, municípios, documentos fiscais, polígonos geográficos e demais objetos de interesse. O problema de associação consiste em determinar se dois registros distintos representam o mesmo imóvel rural.

Definimos a hipótese: $H = \text{"Os registros A e B representam o mesmo imóvel."}$ e a hipótese complementar: $\neg H = \text{"Os registros A e B não representam o mesmo imóvel."}$

3.1 Evidências

Cada agente produz uma função básica de massa:

$m_i : 2^\Theta \rightarrow [0,1]$, onde: $\Theta = \{H, \neg H\}$ e $m_i(H) + m_i(\neg H) + m_i(\Theta) = 1$ sendo:
 $m_i(H)$ = evidência favorável; $m_i(\neg H)$ = evidência contrária; $m_i(\Theta)$ = incerteza.

São definidos quatro agentes: Semântico; Geoespacial; Cadastral e Fiscal produzindo respectivamente: $m_{sem}(H)$, $m_{geo}(H)$, $m_{cad}(H)$ e $m_{fis}(H)$.

Para um par de imóveis A e B , definimos: $X = \{x_1, x_2, \dots, x_n\}$, onde: x_1 = similaridade do nome, x_2 = similaridade dos proprietários, x_3 = coincidência do município, x_4 = coincidência do SNCR, x_5 = coincidência do NIRF, x_6 = similaridade da área, x_7 = sobreposição espacial, x_8 = coincidência do CPF/CNPJ do declarante

3.1.1 Modelagem das Evidências e Confiabilidade das Fontes

Agente Semântico: Compara: Nome do imóvel, Nome da propriedade, Nome da fazenda, Proprietários/Arrendatários. $m_{sem}(H) = Sim_{sem}(A, B)$, com: Similaridade Semântica $Sim_{sem}(A, B) = 1 - Lev(A, B) / \max(|A|, |B|)$, com Lev sendo a Distância de Levenshtein.

CAFIR	SNCR	Similaridade
FAZENDA SANTA MARIA	FAZ SANTA MARIA	0,91

Agente Geoespacial: Compara: Área CAFIR x Área SNCR. $m_{geo}(H) = Sim_{geo}(A, B)$, com: Similaridade de área $Sim_{area}(A, B) = 1 - |AreaA - AreaB| / \max(AreaA, AreaB)$

CAFIR	SNCR	Similaridade
85 hectares	100 hectares	0,85

Agente Cadastral: usa os identificadores. $m_{cad}(H) = Sim_{cad}(A, B)$, com:

$Sim_{cad}(A, B) = (Score_{SNCR} + Score_{NIRF} + Score_{Município}) / 3$, onde $Score_X = \{1, X_a = X_b, 0$, caso contrário, sendo $X = \{SNCR, CAFIR, Município\}$

Mesmo SNCR	Município igual	Similaridade
$m_{cad}(H) = 1$	$m_{cad}(H) = 0,95$	0,95

Agente Fiscal: Agente, Pessoa, Empresa ou Estabelecimento. $m_{fis}(H) = Sim_{fis}(A, B)$ onde $Sim_X = A \times B \rightarrow [0, 1]$

Mesmo CPF/CNPJ	Mesmo Endereço Fiscal	Similaridade
0,79	0,85	0,6715

As evidências são combinadas pela Regra de Dempster. Para duas funções de massa $m_i(H) = Sim_i(A, B)$ e $m_i(\neg H) = (1 - \alpha_i) (1 - Sim_i(A, B))$; $m_i(\Theta) = \alpha_i (1 - Sim_i(A, B))$, onde α_i é a **confiabilidade da fonte de dados**. Por exemplo, considere: $\alpha_{SNCR} = 0,95$, $\alpha_{CAR} = 0,85$.

3. Fusion KG-Rural

A arquitetura proposta é composta por cinco camadas.

Camada 1 — Fontes: CAFIR, SNCR e SICAR/CAR

Camada 2 — Integração Semântica: Conversão dos registros para RDF e OWL.

Camada 3 — Knowledge Graph Rural: Representação integrada das entidades: Imóvel, Proprietário, Município, Polígono e Documento.

Camada 4 — Agentes Especializados:

- **Agente Semântico:** Analisa: nome do imóvel; nomes dos titulares; descrições.
- **Agente Geoespacial:** Analisa: sobreposição de polígonos; distância geográfica; área declarada.
- **Agente Cadastral:** Analisa: NIRF; SNCR; identificadores CAR.
- **Agente Fiscal:** Analisa: declarações ITR; vínculos tributários.

Camada 5 — Fusão de Evidências: Cada agente produz uma massa de evidência: $m_i(H)$. As evidências são combinadas pela Regra de Dempster:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K}$$
, onde K representa o conflito entre evidências. O resultado final produz: crença $Bel(H)$; plausibilidade $Pl(H)$; incerteza $U(H)$.

4. Estudo de Caso

Foram utilizados agentes, para executarem as avaliações independentes para cada par candidato de imóveis. Inicialmente os candidatos foram agrupados por UF. Posteriormente aplicou-se bloqueio por município. Somente imóveis pertencentes ao mesmo município foram comparados. As evidências geradas foram combinadas pelo mecanismo de fusão proposto. Os resultados foram analisados considerando: correspondências confirmadas; correspondências incertas e conflitos entre bases.

Na Figura 1 temos exemplos de pares fortes, mesmo com algumas diferenças entre os nomes das propriedades. Já na Figura 2 temos exemplos dos casos de conflitos, grande divergência entre o nome nas 2 bases de dados.

nome_a	nome_b	nome_levenshtein	nome_token_sort	nome_jaccard	nome_dice	nome_cosine_3gram	nome_semantic_embedding	nome_media_lexical
FAZENDA SERRA DO MACHADO	FAZENDA SERRA DO MACHADO	1	1	1	1	1	1	1
FAZENDA BANDEIRA	FAZENDA BANDEIRA	1	1	1	1	1	1	1
FAZENDA VALE DO GONGOGI	FAZENDA VALE DO GONGOGI	1	1	1	1	1	1	1
FAZENDA CAMPO GRANDE	FAZENDA CAMPO GRANDE	1	1	1	1	1	1	1
FAZENDA AGRO FAMILY	FAZENDA AGROFAMILY	0,95238095	0,952381	0	0	0,800640769		0,541080535
FAZENDA SAO JOSE	FAZENDA SAO JOSE	1	1	1	1	1	1	1
FAZENDA G R	FAZENDA RG	0,4	0,4	0	0	0		0,16
FAZENDA BAHIA	FAZENDA SANTA CACHA	0,375	0,375	0	0	0,104828484		0,170965697
FAZENDA SAO JORGE	FAZENDA SAO JORGE	1	1	1	1	1	1	1
FAZENDA ROSA DE SARON	FAZENDA ROSA DE SARON	1	1	1	1	1	1	1
FAZENDA TRES RIACHO	FAZENDA TRES RIACHOS	0,95652174	0,9565217	0,33333333	0,5	0,815374248		0,712350212

Figura 1. Exemplos de Pares Fortes, com base em função de diversas métricas

nome_a	nome_b	nome_levenshtein	nome_token_sort	nome_jaccard	nome_dice	nome_cosine_3gram	nome_semantic_embedding	nome_media_lexical
FAZENDA BAIXA PRETA II	LAMEIRAO DO COITE	0,322580645	0,193548387	0	0	0		0,103225806
FAZENDA MARIA PRETA	FAZENDA PROGRESSO	0,3	0,3	0	0	0,083624201		0,13672484
NAYARA LINS DINIZ	SITIO PATOS	0,181818182	0,090909091	0	0	0		0,054545455
CHACARA SAO CAETANO	SITIO LAGOA DO EXU	0,347826087	0,347826087	0	0	0		0,139130435
FAZENDA SITIO	FERNANDO SANTIAGO COUTINHO	0	0	0	0	0		0

Figura 2. Exemplos de Casos de Conflitos

Tabela 1. Bel(H) x Qtd.

Faixa Bel(H)	Quantidade
0.95–1.00	89.342
0.90–0.95	67.818
0.85–0.90	42.840
< 0.85	restante

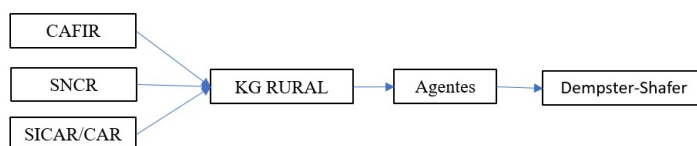


Figura 3. Arquitetura da Solução

5. Resultados

Em termos do SICAR, no Acre foram identificadas 56.398 propriedades, mas só se conseguiu identificar, em relação ao SINCAR, SNCR e CAFIR 13.690 propriedades comuns aos 3 cadastros. Correspondendo a 24,27% do total de propriedades existentes no SINCAR.

Situação	Quantidade	Situação	Quantidade
CAFIR + SNCR	5.081.198	Apenas SNCR	2.837.289
Apenas CAFIR	5.209.557	Total	13.128.044

Observou-se que atributos isolados frequentemente produzem ambiguidades. A combinação de múltiplas fontes de evidência reduziu conflitos e aumentou a confiança das associações.

Os melhores resultados ocorreram quando evidências geoespaciais foram combinadas com evidências cadastrais e fiscais.

Além disso, o uso do Grafo de Conhecimento permitiu rastrear e explicar as evidências utilizadas na tomada de decisão, mas demonstrou que hoje nenhum dos 3 cadastros é capaz de informar todas as propriedades rurais sendo relevante este trabalho.

5.1 Limitações

Os resultados dependem da qualidade dos registros presentes nas bases originais., as quais são incompletas e divergentes. E se tem a inexistência de identificadores universais entre CAFIR, SNCR e SICAR/CAR constitui uma limitação inerente ao problema.

6. Conclusões

Este trabalho apresentou o FusionKG-Rural, uma arquitetura de fusão multimodal baseada em Grafos de Conhecimento e Teoria de Dempster-Shafer para associação automática de imóveis rurais.

Os resultados demonstram que a combinação de evidências heterogêneas constitui uma alternativa promissora para integração de bases territoriais governamentais.

Agente	Evidência	Métrica
Semântico	Nome do imóvel, proprietário	Jaccard, Levenshtein
Geoespacial	Polígono, área, coordenadas	IoU,
Cadastral	SNCR, NIRF, município	Regras determinísticas
Fiscal	CPF, CNPJ, ITR	Similaridade cadastral

Como trabalhos futuros pretende-se incorporar agentes baseados em LLMs locais, Graph RAG e mecanismos de detecção de anomalias em grafos para identificação automática de inconsistências fundiárias e tributárias.

References

- Hall, David L.; Llinas, James (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, v. 85, n. 1, p. 6–23.
- Hogan, Aidan et al. (2021) Knowledge Graphs. ACM Computing Surveys, v. 54, n. 4, p. 1–37.
- Khaleghi, Bahador; Khamis, Alaa; Karray, Fakhri; Razavi, Saiedeh (2013). Multisensor data fusion: A review of the state-of-the-art. Information Fusion, v. 14, n. 1, p. 28–44.
- LI, Yunjia et al. Data Fusion for Knowledge Graph Construction: A Survey. Information Fusion, 2024.
- PAULHEIM, Heiko. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web Journal. 2023.
- Shafer, Glenn. (1976) A Mathematical Theory of Evidence. Princeton: Princeton University Press,
- ZHAO, Zifeng et al. Entity Alignment for Knowledge Graphs: A Survey. IEEE Transactions on Knowledge and Data Engineering, 2023.